How Many Do We Need? Exploration of the Population Size Effect on the Performance of Forensic Speaker Classification

Shunichi Ishihara ¹, Yuko Kinoshita ²

¹ Faculty of Asian Studies, The Australian National University, Australia
² School of Languages and International Studies, University of Canberra, Australia

Shunichi.Ishihara@anu.edu.au, Yuko.Kinoshita@canberra.edu.au

Abstract
This paper investigates how change in population size affects the reliability of the likelihood ratio (LR)-based forensic speaker classification. Using features of the long term F0 distribution, we performed LR-based speaker classification and examined its performance with population sizes from 10 to 120. The results revealed that LRs could be heavily influenced by the population data. We discovered that the reliability of LR-based evaluation of the evidence was heavily compromised if the population data was limited to a small number of speakers.

Index Terms: forensic, speaker classification, population, long term F0 distribution, MVLR, likelihood ratio, spontaneous speech, Japanese

1. Introduction
In the field of forensic science, use of LR-based evaluation is gaining ground. Until recent years the use of LR had been limited to a small number of fields, such as DNA, but the presentation of LRs as a measure of the strength of evidence is becoming a norm especially since the Daubert ruling, which defined criteria for the admissibility of scientific evidence in court.

In order to evaluate the strength of evidence, we need a background population data that represents a relevant population. We thus need a sufficiently large population dataset, with recording conditions comparable to the recordings in question. Though an obvious requirement, in practical forensic casework contexts it is often difficult to achieve, mainly due to time and resource constraints.

Then naturally we must ask ourselves: “How large does the population data need to be for a reliable evaluation of speech evidence?” and, possibly more importantly, “What is the nature of the errors that small population datasets bring?” It is well known that population dataset size affects the accuracy of the classification. However, in order to present evidence in court, experts need to know the characteristics of errors associated with the size of the population data used in the case, as well as its error rate. This paper thus explores the effects of population dataset size on forensic speaker classification, using features extracted from long-term F0 (LTF0) distribution.

2. Experiment
2.1. Use of the long-term distribution of F0
F0 has been a popular parameter in traditional FSR. It has many attractive features for forensic speaker identification practitioners, such as robustness, measurability, and availability [1]. On the other hand, quite apart from the many linguistic uses of F0 — which encodes tone, intonation and stress — many non-linguistic factors are known to affect it, including: state of health, emotional changes, discourse genre, noisiness of the environment, and whether or not the person is on the phone [2], [3], [4]. Thus many (eg. [5], [6]) have noted that a single speaker can show large variation in F0 from occasion to occasion, and even within a single recording session. F0 is therefore considered not very effective as an FSR parameter, and Kinoshita has demonstrated that because of its poor between-speaker variance ratio, mean LTF0 shows very poor strength of evidence, typically generating LRs of effectively unity [7].

However, it has also been reported that other features of LTF0 could provide us with quite useful features to classify speakers [8]. This research revealed that using six features extracted from the overall shape of LTF0 distribution, namely mean, standard deviation, skew, kurtosis, mode, and the probability density at the mode, could classify speakers effectively, achieving equal error rate (EER) of 10.7%.

Following this promising result, this paper uses the six features used in the aforementioned study to examine the effect of the population size on LRs.

2.2. Database and speakers
For this study, we used male Japanese speakers selected from the Corpus of Spontaneous Japanese (CSJ) [9]. CSJ is a database which consists of various styles of speech recorded from 1464 speakers. The majority of the recordings were made in the style of either Academic Presentation Speech (APS) or Simulated Public Speech (SPS). APS was mainly recorded live at academic presentations, most of which were 12-25 minutes long. For SPS, 10-12 minute mock speeches on everyday topics were recorded. In CSJ, all recordings were made using DAT and down-sampled to 16 kHz, with 16 bit accuracy. CSJ incorporates a five-scale evaluation of various aspects of the recordings. We used one of those evaluations, the so-called spontaneity scale, for our speaker selection. By spontaneous, CSJ means ‘sounding as if it is not read out’. (In a situation such as an academic presentation, it is not uncommon for presenters to actually read out their prepared scripts.) In order to simulate forensically realistic conditions, we first selected speakers ranked reasonably highly (three to five) on the 1-5 spontaneity rating. The authors, who are native Japanese speakers, listened to a selection of the highly ranked recordings. They can confirm that they do indeed sound natural and not read out.

Another criterion for speaker selection was the availability of non-contemporaneous recordings. Our spontaneous sounding speakers had to have been recorded on...
two or more different occasions in order for us to attempt a forensically realistic discrimination. On the basis of these two criteria, then, we selected 241 male speakers, with two non-contemporary recordings for each speaker.

2.3. F0 extraction and parameterisation

F0 was extracted using the ESPS routine of the Snack Sound Toolkit [10] with Tcl at every 0.005 second. CSJ usefully annotates non-speech noise with a noise tag. The sections with this noise tag were excluded from the data.

The distributions of the extracted F0 were then parameterised. For each of the 482 recordings, as well as long-term mean and SD, four other parameters which relate to the shapes of the F0 distributions were calculated: skew, kurtosis, modal F0, and the modal density.

Skew measures degree of distribution asymmetry, and kurtosis measures peakedness of a distribution. They are thus useful measures to characterise the overall shapes of the distributions. The mode shows the most commonly occurring F0 for each recording, and modal density represents how concentrated it is.

In order to extract modal F0 and the modal density, firstly the probability density of the sampled F0 for each recording was estimated using binned kernel density (with the \texttt{bkde} function of R statistical package’s KernSmooth library). The appropriate kernel density bandwidth was selected using direct plug-in methodology (the \texttt{dpiik} function of the same library) [11, 12].

2.4. Experimentation process

The experiment in this study took three steps. First of all, we made 12 differently sized population groups, varying from 10 to 120 in increments of 10 speakers. We then selected two groups each from those 12. We thus obtained 24 population groups (2*12 different sizes) as shown in Table 1. The two groups are completely independent from each other, meaning no speaker was included in both groups — $P_{10}^1$ and $P_{10}^2$ do not include same speakers, and likewise, neither do $P_{120}^1$ and $P_{120}^2$. 120 is the maximum size which allows us to create two independent population groups with the 241 speakers.

Table 1. Two groups for each of 12 different population sizes

<table>
<thead>
<tr>
<th></th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>...</th>
<th>100</th>
<th>110</th>
<th>120</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group1</td>
<td>$P_{10}^1$</td>
<td>$P_{10}^2$</td>
<td>$P_{30}^1$</td>
<td>...</td>
<td>$P_{100}^1$</td>
<td>$P_{110}^1$</td>
<td>$P_{120}^1$</td>
</tr>
<tr>
<td>Group2</td>
<td>$P_{10}^2$</td>
<td>$P_{20}^2$</td>
<td>$P_{30}^2$</td>
<td>...</td>
<td>$P_{100}^2$</td>
<td>$P_{110}^2$</td>
<td>$P_{120}^2$</td>
</tr>
</tbody>
</table>

Then we compared pairs of the 241 speakers (speaker pairs), calculating the Multivariate Likelihood Ratio (MVLR) using the formula by Aiken and Lucy [13] against each of the 24 population groups. For each of the 12 different population sizes, we have. Thus, for each speaker pair we produced six logLRs for each of the 12 different population sizes.

Finally, in order to investigate how each comparison is affected by differences in both composition and size of the population group, the range of the two logLRs (logLR range) produced by two population groups was calculated for each comparison (eg $P_{10}^1$ vs $P_{10}^2$). Our hypothesis is that the logLR range should become smaller as the size of population increases.

Setting the three steps described above as one experiment, we performed this experiment three times, each time selecting different sets of speakers for the population groups to obtain a better picture of the situation with the limited amount of data we have. Thus, for each speaker pair we produced six logLRs for each of the 12 different population sizes.

3. Results

In this section, we present and discuss the results of the experiment. The effects of population size differences are examined regarding three aspects: stability, discrimination power, and the accuracy of the evaluation of LR.

3.1. Stability

If the LR calculation is not affected by differences in the composition of background population data, the LRs produced using two different sets of individuals in the population data (eg, the logLRs produced with $P_{10}^1$ and with $P_{10}^2$) should remain very similar to one another. Thus the difference between the two logLRs should reflect the stability of the LR against different selections of the population. Out of the three logLR range values for each comparison, the middle values were pooled and plotted as a function of population size in Figure 1.

![Figure 1](image.png)

Figure 1: Summary of the variability score from three repetitions of the experiments.

The two box plots present separately the results for the two types of speaker pairs: same- and different-speaker. The sizes of the boxes reflect the range of the middle 50% of values. The line in each box indicates the median, and the dot on each box indicates the mean value. We could not calculate the mean for the different-speaker comparisons for population size 10 and 20 due to calculation overflow.

Figure 1 reveals two things. Firstly, population size affects the LR. Secondly, when the population size is small,
the discrepancy between pairs of LRs becomes significant, indicating that the composition of the population has a significant effect. The effect is very large when the population size is 10, and still quite significant with 20. After the population size reaches 30, the effect seems considerably less significant.

The magnitude of the discrepancy between the pairs of logLRs produced with small population data must be noted too. The bottom panel in Figure 1 shows that, for the different-speaker pairs, if we have only 10 people in our population data, the very same speaker pairs can produce enormously different logLRs (in the order of log $10^{13}$) depending on the individuals included in the population. This is an extremely disconcerting result.

3.2. Discriminability

In this section, we examine the effect of the population size on the discriminability. We measured the discriminability in terms of its equal error rate (EER) — where the error rate for the same- and different-speaker comparisons achieve the same error rate — which serves as a convenient shorthand for the discrimination ability of a given system. The summary of the EER is presented in Figure 2.

The x-axis shows the different sizes of the population, and the y-axis shows the EER. For each population size, there are six dots, representing the results from six logLR calculations (two different population groups * three repeats of the experiment). The EER obtained using all 241 speakers (horizontal line at 0.086 = 8.6%) is shown as the benchmark. Firstly, we can see that the EER improves with increase in population size. The improvement seems more rapid up to the population size 30. It continues to improve after this point, but more slowly. By the time 120 people were included in the population data, the EERs become quite close to the benchmark EER. We also can see that this LR-based speaker discrimination is reasonably effective even using very small population data, such as 10. The worst EER obtained was about 17.2%, which is still useful as additional information. In the observation made in the previous section, we found that, however, the actual logLRs are highly variable depending on the set of individuals included in the population. This seems to suggest that we can potentially discriminate same-speaker pairs from different-speaker pairs even when we have a small population size, but we should not trust the logLR we obtain as the indicator of the strength of the evidence.

The mean of the six EERs for each population size are also plotted over this figure in red.

3.3. Calibration

For forensic science practitioners using an LR-based approach, whether or not the LRs obtained as the result of their analyses in fact correspond to the strength of the evidence in reality is a serious concern. The discrimination experiments tell us how often we get LRs which support counter-factual hypotheses, but this kind of binary judgment will not shed any light on the question, such as “does this evidence really have the strength of LR 30 (neither 2 or 1000)?” In case of speech evidence, there is no single parameter that can classify speakers reliably by itself, as speech is an outcome of a complex interaction of many factors. In other words, even if the given two recordings are indeed from the same speaker, some of the parameters used in the analysis could produce an LR which supports the defence hypothesis. Supporting a counter-factual hypothesis by itself is not wrong — it is a part of the nature of an LR, as an LR talks merely about “likelihood”. However, if the methodology produces LRs which do not correspond to the true strength of evidence, combining those LRs could result in extremely misleading evaluations of evidence.

In order to address this problem, Brümmer and De Preez introduced the idea of assessing “goodness of calibration” [16] [17]. In this approach, the “goodness” of the LR is analysed in detail, using techniques such as calibration- and discrimination-loss decomposition of the
results, with associated Applied Probability of Error (APE) plots. This would greatly assist our understanding of the results, and we intend to expand our investigation using these tools in the future. In the current preliminary study we concentrate our analysis on a much coarser indicator of the calibration of the test — the location of the EER in relation to logLR 0.

![Figure 3: Summary of the distance between logLR 0 and EER calculated for differently sized population data.](image)

Aitkin and Lucy’s MVLR formula produces generative LRs, meaning that if the piece of evidence is information neutral, the logLR should be 0. We thus expect that if this testing system is working well the EER should fall in the vicinity of 0. In this study, we thus looked at the discrepancy between logLR 0 and the logLR where the test achieved EER.

The results are summarised in figure 4. The x-axis shows the population size, and the y-axis shows logLR where the test achieved EER. The horizontal line at logLR 0 would be an ideal system, and distance from this represents inaccuracy in the calibration of the system. As with the figure in the previous section, each population size has six points, representing the results from six different calculations. The mean of the six calculations for each population size is also plotted as a solid line. The figure reveals a clear tendency that the calibration of the test improves as we increase the size of the population data and there is a particularly significant improvement between the population size 10 and 20. Furthermore, the size of the discrepancy seems to suggest that we really should not be relying on LRs produced using smaller population data. With the population size 10, the LR test produced an EER at logLR -6.02 on average. In other words, the information neutral point with this test fell at -4.02, far off from 0. As a logLR of -4 is generally considered to be very strong support for the defence hypothesis [18], this indicates that such a result is extremely poorly calibrated.

4. Discussion and future direction

In this paper, we examined how population size affects MVLR based forensic speaker classification. Our investigation focused on three aspects: stability, discriminability, and calibration of the testing method. We discovered that the population size does have a significant effect on the reliability of LRs. In all three aspects, we had very similar results. There was a clear tendency that the performance of the test improved as we increased the size of the population; when we increased the population size from 10 to 30, the improvement was very significant.

The average EER of 14.7% that we obtained from the test with population size 10 seems to indicate that even when we have a very small population data, this approach may have useful discriminating power. This is welcome news, however, we also found that the actual logLR of this testing will be totally unreliable from observation of its stability and calibration.

In sum, the results of this preliminary study seem to suggest that we do need a large population data in order to produce reliable MVLR; and LR produced using anything smaller than 30 people as its population data is highly unreliable.

As a future task, we plan to explore the option of calibrating the MVLR results, using De Preez and Brümmer’s calibration technique. This would allow us to shed more light as to how the population affects the MVLR. Also, if we could adjust calibration, we may be able to transform the currently highly unreliable LRs produced with a small population dataset into something more reliable. It would make an extremely beneficial contribution to the field of forensic speaker classification.

5. References